The Impact of After-School Programs: Interpreting the Results of Four Recent Evaluations

by

Thomas J. Kane
University of California, Los Angeles

January 16, 2004

This is a working paper of the William T. Grant Foundation and should be cited as such.

For Immediate Release: January 16, 2004

The following report was written with support from the William T. Grant Foundation. Robert Granger, president of the Foundation, requested the review and provided a number of insightful comments along the way. The authors of the reports that are reviewed—Mark Dynarski, Jean Grossman, Elizabeth Reisner, Karen Walker and Richard White—provided comments on earlier drafts. A number of influential scholars, advocates, policymakers, and practitioners—including Phoebe Cottingham, David DuBois, Samuel Halperin, David Hansen, Jackie Kaye, Peter Kleinbard, Joseph Mahoney, Edward Pauly, Karen Pittman, Stephanie Schaefer, Allen Schirm, Edward Seidman, Christine Sturgis, Grover Whitehurst and Nicole Yohalem—provided comments on the penultimate draft. Susan Teitz from the William T. Grant Foundation provided much-needed editorial help. Marina Bassi provided assistance in collecting data on Stanford 9 scores across grades. Despite their diligence, any remaining errors remain the author's responsibility. Any comments or questions should be submitted to the author at tomkane@ucla.edu.

I. Introduction

Within the last decade, after-school programs have moved from the periphery to the center of the national education policy debate. The demand for after-school care by working parents and a new focus on test-based accountability are the two primary reasons. Reflecting these pressures, federal funding for after-school programs has grown dramatically over the last half-decade. Between 1998 and 2002, federal funding for the 21st Century Community Learning Centers program grew from $40 million to $1 billion. State and local governments have also increased their funding, with California committing itself to a six-fold increase in funding for after-school programs over the next few years.

As a wave of evaluation results has recently become available, policymakers are understandably eager to see evidence that these investments are paying off. The purpose of this review is to summarize the results of four recent evaluations, to draw the lessons we have learned so far, and to identify the unanswered questions. The following four studies are discussed (each of the following evaluations is described in more detail in Table 1):

21st Century Community Learning Centers (21st CCLC): Conducted by Mathematica Policy Research and Decision Information Resources and released in January 2003, the evaluation focused on a sample of elementary and middle school centers. The middle school evaluation used statistical controls to compare participants and non-participants attending the schools where after-school programs were located. The evaluation of elementary school programs used random assignment, for a sample of 18 centers that had more applicants than slots. (This sample of centers was expanded to 34 centers in a second year of results, not yet released.) The evaluation is ongoing.

The After-School Corporation (TASC): Conducted by Policy Studies Associates, the evaluation focuses on 96 after-school centers in New York City over four years. The evaluation did not use random assignment, but used statistical controls to compare outcomes for participants and non-participants. Both groups were enrolled in the same schools where the after-school programs were located. The evaluation is continuing. Data collection is completed and the final analyses and report are being prepared.

Extended-Service Schools Initiative (ESS): Conducted by Public/Private Ventures and MDRC, the evaluation focused on ten after-school centers in six cities. The evaluation design was non-experimental. Unlike the 21st CCLC and TASC evaluations, the ESS evaluation focused on youth participation and cost and was not designed to identify impacts on other outcomes, such as student test scores or grades in school. The evaluation is completed.

San Francisco Beacons Initiative (SFBI): Conducted by Public/Private Ventures, the evaluation focused on five after-school centers in San Francisco. The evaluation design was quasi-experimental, comparing program participants with non-participants. The evaluation collected information on the nature of activities at the after-school programs and rates of participation, as well as test scores and school grades. The evaluation is completed.

While the remainder of this article provides a more detailed discussion of the findings and methods used in each of the evaluations, four points deserve emphasis:

First, among centers that operated on a drop-in basis, attendance was sporadic. The typical participant attended one to two days in the average week. Given the sporadic attendance, one would expect only modest impacts on outcomes such as performance on achievement tests, effort while in school, and parental participation for the average participant.

The TASC program in New York City proved the exception, where the average elementary school participant attended 3.9 days per week (and middle school students attended 2.9 days per week). Understanding the source of the difference in attendance rates should be a high priority.

Second, the 21st CCLC evaluation failed to find a large impact on the proportion of children without supervision after school. Traditionally, the primary purpose of after-school programs has been to increase adult care during after-school hours for children whose parents would otherwise be at work. Despite the stated rationale, it is regrettable that only one of the evaluations—the 21st CCLC evaluation—attempted to compare the after-school care arrangements of participants and non-participants. Perhaps surprisingly, the evaluation suggested that few of the participating children would have been at home alone without the program. In the 21st CCLC middle school evaluation, participants were 8.5 percentage points more likely to report being cared for by a non-parental adult after school (such as an after-school care staff member). But they were also 6 percentage points less likely to report being cared for by a parent and 2.6 percentage points less likely to report being cared for by a sibling. (They were no more likely to report being alone after school.) In other words, the increase in non-parental adult care in the after-school programs was offset by a decline in parental care and sibling care. Moreover, two-thirds of the net increase in non-parental adult care (6 percentage points out of the 8.5 percentage points) seems to have come as a result of a decline in parental care. (The remaining third was due to a decline in sibling care.)

Because the 21st CCLC middle school evaluation relied on statistical controls to account for the differences between the participants and non-participants, the middle school results may be biased against finding an impact on after-school care—since non-participants may have had better alternative care options at the baseline than participants. However, the elementary school evaluation—which used a random assignment design— reported similar results. The 10.5 percentage point increase in the proportion of youth receiving non-parent adult care after school was offset by a decline in the proportion of youth being cared for by a parent (7.4 percentage points) or sibling (2.7 percentage points) after school. There was no difference in the proportion of parents reporting that the children were caring for themselves after school.[1]

[1] Unfortunately, the measure of after-school care used in the 21st CCLC evaluation suffered from a number of weaknesses, described in more detail in this report. Such weaknesses may have led the evaluators to overlook small or moderate impacts on after-school care. However, there is no evidence that the 21st CCLC programs had large impacts on the proportion of youth caring for themselves after school.

As described later in this report, there are several problems with the measure of after-school care used in the 21st CCLC evaluation that may have led to an understatement of the impact. (Given these results, future evaluations should focus more resources on measuring such impacts reliably.) Nevertheless, the failure to find a large impact on the proportion of children in self-care has important implications for interpretation. If the programs had been shown to substantially reduce the proportion of children caring for themselves, many observers may have been willing to grant such programs the benefit of the doubt with regard to other outcomes. Impacts on outcomes such as academic achievement or parental engagement in a child's education may be difficult to discern, even if they are sufficiently large to be worthwhile. (This is particularly likely to happen when participation rates are below those expected by the evaluators.) Even in the absence of statistically significant impacts, readers may have been more willing to presume that structured activities in a safe environment would have beneficial impacts for children—if those children would otherwise be caring for themselves alone. If the children would not have been unsupervised after school, parents and policymakers will need to be persuaded that a couple of hours in an after-school program is more worthwhile than the same time spent at home. That may well be true—given that some home environments are not ideal—but the burden of proof is likely to be higher.

Third, none of the evaluations reported a statistically significant impact on achievement test scores after one year of participation. The TASC evaluation failed to find impacts on math or reading achievement test scores in the first year, but did report positive impacts on math performance after two and three years of participation. Unfortunately, the non-experimental design of that evaluation and the potential unmeasured differences between non-participants and those who remain in the program over the long term casts some doubt on these estimated long-term impacts. Interestingly, even among the long-term, active participants, there was no evidence of reading score impacts in the TASC evaluation. The SFBI evaluation—another non-experimental evaluation using statistical controls to compare participants and non-participants—also failed to find impacts on grades, test scores, or school absences. The only evaluation using random assignment—the 21st CCLC elementary school evaluation—failed to find impacts on reading scores. (Unfortunately, that evaluation did not collect follow-up math scores that would have allowed for a comparison with the TASC evaluation's more positive results.)

However, we should probably not be surprised by the failure to find statistically significant impacts on achievement tests in most studies given the lack of statistical power. The 21st CCLC evaluation was designed to be able to identify a .20 standard deviation impact on reading test scores.[2] As argued below, this is an unrealistically large impact given the nature of the intervention. In the national samples used to norm the Stanford 9 test, students taking the test in the spring of fifth grade scored approximately one-third of a standard deviation higher in reading and one-half a standard deviation higher in math than students taking the test in the spring of fourth grade.[3] In other words, everything that happens to a student between the end of fourth grade and the end of fifth grade—a whole school year of full-day classroom instruction

[2] With test score outcomes for 285 treatment group members and 156 control group members, the evaluation had about a two-thirds chance of rejecting a .20 standard deviation impact in a two-tailed difference of means test at the .05 level.

[3] Harcourt Educational Measurement (1996), Tables N2 and N5.

and homework, a school year and a summer of conversations with parents and family at the dinner table—is associated with one-third to one-half of a student-level standard deviation gain in performance. In light of these gains, a .20 standard deviation gain based upon an additional hour of academic instruction per day would have been quite large—even if the programs achieved 100 percent participation.

Unfortunately, it has been common practice in evaluation research to design studies to detect a .10, .20, or even .30 standard deviation impact, regardless of the nature of the intervention. But, clearly, the size of the impact one might reasonably expect should be a function of the nature of the program being evaluated.

I present two ways to think about setting a more reasonable expectation for educational interventions such as after-school programs. One approach is to start with an estimate of the impact of a given amount of classroom instruction on academic performance. For example, if we had an estimate of the impact of the standard 180-day school year on achievement test performance, we could generate an estimate of the expected gain from spending a fraction of that time in an after-school program. The actual impacts may be larger (if the program consolidates the gains from regular course instruction by helping students learn from their homework) or smaller (since students are tired at the end of a day). However, the impact per hour of regular course instruction is a reasonable place to start in forming baseline expectations of after-school programs.

Alternatively, we could study the relationship between test performance and earnings later in life and calculate the magnitude of the test impact that would generate an increase in later earnings sufficient to cover the cost of the intervention. In the case of after-school programs, both approaches would lead one to expect an impact between .05 and .07 student-level standard deviations—even with 100 percent attendance at the after-school programs. In other words, the magnitude of the impacts that would reasonably be expected is one-quarter to one-third the size of the impacts that at least one evaluation was designed to detect.

Unfortunately, all such reasonably sized impacts lie within the confidence intervals (or bounds of statistical precision) reported for the 21st CCLC evaluation. Because those confidence intervals on the achievement score gains also include zero, the results are often interpreted as showing that the program had no effect on academic performance. But, with such statistical imprecision, failing to reject the hypothesis of no impact need not imply that the programs are having no effect. Rather, we may have only learned that the impacts are not extraordinarily larger than the effect of the average hour of classroom instruction during a school year or extraordinarily larger than would be necessary to pay off the cost of the after-school programs.

In designing future evaluations and in interpreting the results of existing evaluations, we need to be more systematic in thinking about the likely achievement test impacts. In doing so, we may discover that the sample sizes required to identify the expected impacts on academic achievement would be prohibitively expensive. In those cases, identifying intermediate outcomes on the road to student achievement—including parental involvement and homework completion, as well as other outcomes, such as teacher perceptions of student engagement—may be all we can expect.

Fourth, there are fairly consistent results across several of the evaluations suggesting that after-school programs promote greater parental involvement in school, greater student engagement, and greater student commitment to homework. Each of the reports includes findings that the parents of participants were more likely to attend parent-teacher organization meetings, after-school events, and open houses. Even the random assignment evaluation of 21st CCLC elementary school programs reported higher parental attendance at school events and increased parental help with homework for those participating in the after-school care programs.

Moreover, although there were no statistically significant impacts on achievement test scores after one year of participation, there is some evidence that participants did their homework more consistently and, in some cases, achieved higher grades in school.

In the remainder of this article, I first briefly describe the nature of the programs that have been evaluated. (For a more detailed description of the programs, the interested reader is encouraged to study the reports themselves.) Subsequent sections discuss the participation rates observed in each of the programs, the design of the evaluations, and the impact estimates themselves. A final section discusses the magnitude of achievement test impacts that might reasonably be expected.

II. Brief Description of the Programs

Although not all of the programs followed the standard model, the typical after-school program included in these evaluations operated for two to three hours at the end of the regular school day, four to five days per week. The first hour was typically devoted to academic content, although the degree of adult supervision and the mix between instruction and simple homework completion varied. For example, students in some programs were expected to work independently on their homework. In other programs, students would break into small groups to work with staff, receiving additional instruction related to the homework content. The hour of academic content was typically followed by an hour of organized activity involving games, athletic activities, presentations by local community groups, or training in personal skills such as leadership or conflict resolution. Student to staff ratios were typically about 11 to 1.

The after-school programs were typically located in neighborhood schools serving a high concentration of disadvantaged students. For example, in the sites funded by TASC in New York City in 2000-01, 81 percent of students at the schools (participants and non-participants) were African-American or Hispanic and 88 percent were eligible for the free or reduced-price lunch program. In the ESS evaluation, 52 percent of youth participants were African-American or Hispanic and 72 percent participated in the free or reduced-price lunch program. The youth served under the SFBI programs were less disadvantaged by these measures. In the SFBI programs, 50 percent of the participants (and 28 percent of the non-participants at the host schools) were free or reduced-price lunch recipients and 78 percent were African-American,

Hispanic, or Asian/Pacific Islander (of whom the vast majority, 48 percentage points, were Asian/Pacific Islander).

Program Cost

In the 21st CCLC evaluation, the typical middle school program had a student-to-staff ratio of approximately 11 to 1 and staff members received an hourly wage of $16 per hour. This would have implied a cost of approximately $5 per youth slot per day for staff members. Administrative support would have added to the cost.

The ESS evaluation reported a cost of $15 per youth slot per day, although this ranged from $8 to $36 across the ten sites in that study. This figure includes the direct costs of the program, as well as the costs born by others providing services to the children while they were in school. However, the figure excludes the cost of building space and utilities. With the average enrollee attending 40 days per year (20 per semester), the cost per enrollee was roughly $600 per year.

The SFBI evaluation reported average costs of $27 per youth slot per day, although the costs varied widely (from $15 to $41 per youth slot per day) at the five centers. These costs included matching funds provided by other organizations.

The TASC evaluation reported costs of $6.76 per youth slot per day—considerably lower than the cost of the other programs. While this excluded building costs and utilities, like the other estimates, it also excluded some services funded outside the TASC grant. If these other costs had been included, the cost of the program may have been more similar to the others.

III. Rates of Participation

As summarized in Table 2, each of the programs reported participation differently, making it difficult to make direct comparisons. For example, the 21st CCLC middle school evaluation defined a "participant" as a youth who attended a center at least three times during the initial month of operation. In the ESS and SFBI evaluations, "participation" required attendance in at least one session during a school year. The TASC evaluation reported the proportion of youth ever participating. However, much of the TASC evaluation focuses on "active participants"—those who participated in the program at least 60 days over the school year and at least 60 percent of the time while they were enrolled.

Whatever the definition used, the typical participant seemed to attend very sporadically. The average days attended by middle school participants was .9 days per week in the 21st CCLC evaluation and 1.2 days per week in the ESS evaluation. Participation was slightly more regular in elementary schools, 1.9 and 2.4 days per week in the 21st CCLC and ESS evaluations, respectively.

Only the TASC evaluation reported considerably higher attendance rates: 2.9 days per week for the average middle school participant and 3.9 days per week for the average elementary school participant. There are at least three possible explanations: First, the program asked parents to commit to leaving their children five days per week, rather than operating on a drop-in basis. [4] The need for an upfront commitment may have led parents to drop off their children more regularly (or may have led those with less need for regular after-school care to self-select out of the program). However, because families were rarely dropped because of sporadic attendance, and because the cost of the penalty—being dropped from the program—would have been least costly precisely for those expecting sporadic attendance, this explanation seems unlikely to account for all of the difference. Second, the schools served under the TASC program were particularly high-need schools—88 percent of the children in the site schools qualified for the free or reduced-price lunch program. However, the elementary schools in the 21st CCLC evaluation were also high-need schools—three-quarters of the centers were in schools with more than 75 percent free and reduced-price lunch eligibility—but their participation rates were substantially lower than achieved under the TASC program. A third possible explanation is that the New York City Department of Youth and Community Development based its contribution to the program on actual enrollment using audited enrollment reports. This financial incentive may have led programs to find ways to raise attendance.

The evaluations also seemed to differ substantially in the proportion of eligible youth ever participating in after-school programs. Two of the evaluations reported high proportions of eligible youth participating in after-school programs at some point over the year. For instance, 47 percent of the children who were enrolled in the middle schools that served as hosts in the SFBI programs attended the after-school programs at some point. The programs included in the ESS evaluation reported more than 50 percent of youth from the host schools participated at some point. The TASC evaluation reported 39 percent of youth from the host schools as ever participating.

In contrast, among those completing the baseline survey in the 21st CCLC evaluation, only 12 percent had participated in the after-school program.[5] The lower participation rates were at least partly due to the fact that participants were required to have attended three or more sessions during the first month of the school year to qualify. Since many of the programs were new, some parents may not have known about them in time to participate within the first month. Presumably, over the course of a school year, the total number of participants would have been much greater.

IV. Evaluation Designs

The primary challenge for any impact evaluation is the construction of a plausible estimate of what would have happened to the treatment group in the absence of the program

[4] Three of the ten sites described in the ESS evaluation also had mandatory attendance policies; these sites also had higher participation rates. Grossman et. al., pp. 15-16.

[5] The appendix reports the proportion of survey respondents participating (p. 114). Presumably, this may be different from the percent of all youth enrolled in participating schools.

being evaluated. Only the evaluation of 21st CCLC elementary school centers used random assignment to do this. (Because the elementary after-school centers were over-subscribed, the evaluation team was able to randomly choose treatment group members from among a pool of applicants.) The remaining evaluations, including the 21st CCLC middle school evaluation, used statistical methods, controlling for observed differences between those who participated in the after-school programs and a comparison group. The comparison group in the non-experimental studies consisted of non-participants *attending the same schools*, who for some reason did not participate in the programs.[6]

In the evaluation of middle school programs, the 21st CCLC evaluation initially used propensity score matching to identify a comparison group among the non-participants. This was done using the data collected on the student baseline survey data. Because they had not been collected at the time of the propensity score matching, the parent survey data and much of the school administrative data were not used when doing the propensity score matching. The matching was done using only those characteristics available on the student baseline questionnaire—race, gender, grade, student-reported average grades, homework, the reported number of hours reading and watching television, student expectations of dropping out of school, student-reported parental education, and a number of other indices of student responses—and did not include parent information, prior after-school care usage, or baseline test scores.[7]

When the research team collected additional data for the treatment and comparison groups, the groups differed on these other characteristics (in some cases dramatically). The comparison group had statistically significantly higher grades, watched less television (2.01 hours per day versus 2.14 hours), had higher parental education (32.5 percent with at least one college graduate parent versus 29 percent), and higher family incomes (19.6 percent with incomes over $60,000 versus 14.0 percent) than the participants.

For the impact estimates, the analysts in the 21st CCLC middle school evaluation were forced to resort to regression-based approaches to control for these other characteristics. Once the large difference in baseline characteristics between treatment and control groups was discovered, the evaluators were in the same position as the other non-experimental studies, in having to assume that they were controlling appropriately for the many relevant ways in which the treatment and control groups differed.

As reported in the 21st CCLC middle school evaluation, the participants tended to come from more disadvantaged backgrounds than the comparison group of non-participants.[8] In the SFBI evaluation, participants had lower baseline GPA's and achievement test scores than non-participants. Only in the TASC evaluation were the baseline characteristics of participants and non-participants similar on most dimensions—including race, free or reduced-price lunch eligibility, immigrant status, and initial math and reading achievement.

---

[6] The 21st CCLC elementary school evaluation also used regression adjustment. However, this was primarily to increase the precision of the estimates, rather than control for pre-existing differences between the treatment and control groups.

[7] For a complete list, see Table B.1 of the 21st CCLC report.

[8] Table A.8 of the 21st CCLC report.

Because participants chose to attend the after-school program, it is reasonable to presume that on such things as feeling of safety after school and availability of alternative after-school arrangements, the treatment and control groups were, in fact, quite different. (Why else would the participants attend, but seemingly similar non-participants not attend?) As a result, in the quasi-experimental evaluations, we need to be cautious about interpreting impacts on those outcomes most likely to be related to the decision to participate. For instance, one must be careful in interpreting the impacts on after-school care arrangements and feelings of safety after school, given the strong presumption that parents choosing to send their children to an after-school program probably had fewer alternative after-school care options and, possibly, felt less comfortable having their children return home after school.

In some cases, the observable differences between participants and non-participants on characteristics such as race and parental education were small. However, this is not necessarily a source of comfort. Given the very different behaviors of participants and non-participants in showing up for after-school care, they clearly differ along *some* dimension. As discovered in the 21st CCLC middle school evaluation, participants and non-participants who were matched to be similar on an initial set of variables were subsequently revealed to be quite different when additional measures from the parent and school databases became available. We simply may not be measuring the relevant background characteristics.

Rather than simply listing the prior characteristics on which the treatment and control groups differed, a more compelling test would be to see if the regression adjustments are sufficient to eliminate any differences in the key outcomes in the baseline—before the intervention. For example, if one collected data on the ultimate outcomes of interest at the baseline (e.g., after-school program participation, test scores, grades, homework completion, and parental engagement with the schools), one could test whether the regressors were sufficient to eliminate any prior difference between the treatment and comparison groups on these key outcomes before the program existed. The 21st CCLC middle school evaluation performed such an analysis, at least on a subset of the key outcomes. After regression-adjustment, statistically significant differences in baseline grades were reduced to statistical insignificance.[9] However, statistically significant differences in indices of homework habits and disciplinary problems at the baseline remained even after regression-adjustment.

Baseline test scores were available only for a small subset of youth in the middle school sample (about 36 percent) in the 21st CCLC evaluation. For this subsample, the baseline difference in reading and math scores was reduced by one-third, but remained statistically significant after using the other regressors to "adjust" the difference between participants and the comparison group.

The presence of such differences in the baseline characteristics in the 21st CCLC middle school evaluation, even after regression-adjustment, is cause for some concern. However, such concern should not be limited to the 21st CCLC middle school evaluation. Indeed, the 21st CCLC evaluation used a richer set of baseline regressors than some of the other evaluations. Presumably, if similar data were available for the other quasi-experimental evaluations, they would have found similar results.

[9] Based upon correspondence with report authors. Results not included in first-year report.

Response Rates

Compounding the selection problems associated with participation, response rates in baseline and follow-up surveys were often quite low. In addition, baseline data on test scores in administrative records were often incomplete.

ESS: Of the 1708 enrollees who attended at least one day of ESS, 1144 were in fourth to eighth grade. Of these, 69 percent (786) completed a baseline survey. In spring 2001, when the follow-up survey was administered, 674 students completed it. But this included respondents who had not completed a baseline survey. Less than half (371) of those completing a baseline survey (786) completed the follow-up survey. The parent survey was collected from 221 of 336 parents.

TASC: The TASC evaluation relied primarily on school records in its impact estimates. Student test score and attendance data were available for 88 percent of participants and 92 percent of non-participants in 2001-02. The study does report impact estimates for "active participants" and those who participated over several years. However, given the high attrition rates across years and the fact that one-third of participants were "non-active," I will focus on the single-year impacts for all participants.

SFBI: Response rates on the baseline survey varied from 52 to 80 percent in the various schools. However, on the follow-up survey 18 months later, 80 percent of those with a baseline survey responded.

21st CCLC: The response rates in the middle school evaluation follow-up were fairly high: 95 percent, 85 percent, and 78 percent of the students, parents, and teachers, respectively, included in the baseline responded to the follow-up. However, the response rates for the elementary school evaluation were disappointing. For example, the response rates in the follow-up survey of elementary school parents were only 68 percent for treatment and 59 percent for control. For the teacher survey, the response rates were 79 percent for the treatment and 66 percent for the control. The low response rates for the elementary sample are most disconcerting, since attrition may have compromised the comparability of the participants and non-participants achieved with the initial random assignment.

In addition, in the 21st CCLC evaluation, baseline test scores were available for only a small share of the students—73 percent of elementary students and 64 percent of middle school students were missing baseline scores. The school districts were not testing in every grade and the researchers did not have the funds to administer baseline tests. Given the quasi-experimental nature of the evaluation, such data would have been quite helpful. In future evaluations, researchers would be well-advised to plan to collect their own test score outcomes, unless more complete data are known to be available from the school system.

Assessing Some Specific Criticisms of the 21st CCLC Evaluation

Soon after the first-year evaluation of the 21st CCLC program was released, the Bush

administration proposed a 40 percent cut in funding for the program. Reflecting the high stakes involved in the debate, the evaluation has been the subject of considerable controversy since it was released in January 2003.[10] Beyond the issues raised above, questions have been raised about whether the elementary program sites were representative of sites elsewhere, the maturity of the programs being evaluated, and the importance of "cross-over" between the treatment and control groups. I discuss each briefly below:

1. Representativeness of the Elementary School Sites

The middle school evaluation drew a sample of sites that seemed to be representative of the national program. However, for the elementary programs, the random assignment evaluation design forced the evaluators to choose from among the subset of sites with an excess of applicants. As a result, that evaluation was limited to the subset of programs that were oversubscribed. It is uncertain if they were oversubscribed because the programs themselves were of particularly high quality or if those sites were lacking in alternative programs for children. However, the children at the elementary school program sites that were selected for the evaluation were quite different from children at elementary program sites nationally. Sixty-seven percent of the students at the elementary school sites were African-American, as compared with 23 percent of the elementary school sites nationally. More than two-thirds (71 percent) of the elementary school centers were in schools where more than 75 percent of the students were free or reduced-price lunch eligible, compared with 45 percent of the elementary sites nationally.

The critics are right: the elementary sites were not representative of the sites nationally. (Mahoney and Zigler [2003], Bissell et. al. [2003]) But it is worth asking how we might expect this to have affected the results. The programs were non-representative because they were oversubscribed. However, to the extent that the programs were oversubscribed *because* there were few other options available or *because* they were of particularly high quality, we might have expected the impacts at these sites to overstate, rather than understate, the impacts one would have found with a nationally representative sample.

2. Lack of Program Maturity

Critics have also emphasized that the programs included in the evaluation were relatively new and may not have moved beyond the start-up phase to the point where they were implementing the services in the manner intended. (Mahoney and Zigler [2003], Bissell et. al. [2003]) If true, this may have led to an understatement of the impacts achieved by a mature program. All of the programs included in the study had completed at least one year of operation under a 21st CCLC grant and some had completed two years of operation before the baseline data were collected. Moreover, 65 percent of the middle school grantees and 57 percent of the elementary school grantees had operated after-school programs in the schools before they received the 21st CCLC grant. (Dynarski et. al. [2003]) The authors failed to find differences in program impacts for programs that had received grants one and two years earlier.[11] Still, it is

[10] See, for example, Jacobson (2003).

[11] Based upon correspondence with report authors. Results not included in first-year report.

difficult to know the extent to which program maturity may have affected the results. One way to resolve the question would be to see if the programs changed in observable ways in the second year of the evaluation or if the impacts changed. The second year evaluation may shed some light on these questions.

3. Cross-Over Effects

Some of those assigned to the control group in the 21st CCLC evaluation "crossed-over" and participated in the after-school program. (Mahoney and Zigler [2003], Bissell et. al. [2003]) Among students originally assigned to the comparison group, 8 percent of elementary school students and 14 percent of middle school students participated in 21st CCLC programs. (Dynarski et. al. [2003], p. 123) Such cross-over is not extraordinarily large or unusual for this type of study.[12]

Potentially more important for the interpretation of the results is the possibility that comparison group members were participating in after-school programs elsewhere. For example, the elementary school evaluation is only identifying the incremental impact of the offer of participation in the program *at the youths' school*. Accordingly, the impacts will not reflect the value of all after-school program options. Moreover, to the extent that the other programs available to the comparison group may also receive 21st CCLC program support, that incremental impact is not necessarily measuring the incremental impact of the 21st CCLC program as a whole. The impact evaluation is designed to capture only the impact of the program at the child's school.

V. Impact Estimates

A primary rationale for after-school programs is to provide youth with a safe venue for spending their after-school hours, and to encourage them to use the time productively. For example, advocates of Proposition 49 in California (the passage of which will lead to a six-fold increase in state funding for after-school programs in 2004-05) often noted that most juvenile crime offenses occur during after-school hours. The following passage was drawn from a web site sponsored by supporters of the initiative:

> *Q. Do juvenile crime rates really increase after the school bell rings and before parents return home from work?*
>
> *A. Yes, and dramatically so. According to law enforcement data, 3:00 p.m. to 6:00 p.m. is the "prime time for juvenile crime." In the hours between 2:00 p.m. and 4:00 p.m., the violent juvenile crime rate doubles. More than one million K-9 students have no place to*

---

[12] Mahoney and Zigler (2003) report that a large share of those participating in the first year in two sites did not persist in the program for a second year and that a significant number of the non-participants from the first year became participants in the second year. However, this would only be a problem for estimating the impact of two years of program participation. The estimates in the first-year report are unaffected by any subsequent cross-over.

go after school other than to an empty house. Studies show that students who participate in after-school programs are less likely to commit violent crimes, be a victim of a violent crime, skip school and use drugs, tobacco and alcohol. As a result, neighborhoods become safer for everyone. -www.joinarnold.com

By providing a subsidized source of after-school care, the after-school programs are intended to increase the number of children under adult supervision after school and to reduce the number of "latch-key" children. However, not all of the children attending after-school programs would have been wandering the streets without adult supervision in the absence of the program. Some portion of the children presumably would have been at home with a family member or attending some other form of after-school care or activity.

Unfortunately, only the 21st CCLC evaluation provided estimates of the impact of the program on after-school care arrangements. Given the central role of after-school safety and adult supervision in the rationale for after-school programs, this is somewhat surprising. Moreover, both the elementary and middle school components of the 21st CCLC evaluation suggested that a large portion of the attendees would have been cared for at home by a parent in the absence of an after-school program. In the middle school evaluation, participants were 6 percentage points less likely to report being cared for by a parent after school and 2.6 percentage points less likely to report being cared for by a sibling. The increase in care by a non-parental adult was close to exactly offsetting those numbers—a 8.5 percentage point increase. Strikingly, there was no statistically significant change in the percentage of children reporting self-care.

As noted above, the 21st CCLC middle school evaluation relied on statistical controls to create its comparison groups, rather than random assignment. Consequently, we might be concerned that such results would be biased against finding an effect because the participants may have had fewer alternative care options than non-participants. However, the elementary school evaluation—which used random assignment—reported similar results. Parents of participants were 10.5 percentage points more likely to report that their children were receiving non-parent adult care after school. However, they were also less likely to report their children being cared for by a parent (7.4 percentage points) or sibling (2.7 percentage points) after school. There was no difference between participants and non-participants in the proportion of children reporting that they cared for themselves after school.

Given the objective of the program, these findings are potentially quite important. If most participants would have been cared for by a parent in the absence of after-school care, this might explain the failure to find impacts on other important outcomes.[13]

However, the estimated impacts may also have been muted by the nature of the measures

[13] In the 21st CCLC evaluations, neither the elementary nor the middle school evaluations reported statistically significant impacts on the proportion of youth reporting that they felt safe "after school until 6 p.m." As noted above, we might be cautious about the middle school estimates, since one might expect the availability of safe after-school care options to have been negatively related to the decision to participate. The available regressors may not have adequately controlled for the ways in which the treatment and control groups differed. The failure to find an effect for the elementary school children is much more important because of the randomized control design used in that evaluation. However, this may also be understated if students were having difficulty with the distinction between time spent after regular school and time spent after the after-school program.

used. The 21st CCLC middle school students—both treatment and comparison groups—were asked to report "Who was with you after regular school ended until 6 o'clock?" for each day during the week preceding the follow-up survey. (The parents of elementary school students were asked similar questions: "Who was with your child after regular school ended until 6 o'clock?") The available responses were "My mom or dad," "An adult who is not my parent," "An older brother or sister," "A friend who is about my age," "I was by myself," or "Someone else." Respondents were encouraged to "Check all that are true about you." The evaluation then reported the share of participants and non-participants reporting each category at least three days on the prior week. Those who were not in any one category for at least three days were categorized as being in "mixed care."

There are four potential problems with inferring that the program had no impact on after-school care using this measure. First, since the after-school programs were generally conducted on school grounds, respondents may have misunderstood the distinction between time spent in regular school and time spent at the school in the after-school program. The middle school student follow-up question asked students to report how they spent their time "after regular school ended until 6 o'clock." The typical program ended *before* 6 p.m. If the care arrangement immediately after regular school was similar to the care arrangement between 5 p.m. and 6 p.m., the measure would tend to find no difference if students had difficulty with the distinction. In other words, the measure would not capture any reduction in the *amount of time* spent at home alone, even if there were such an impact.

Moreover, any student who reported care by a parent before 6 p.m. was counted as being in a parent's care after school—even if they had been alone from 3 to 5 p.m. This could also lead to an understatement of the impact. For example, suppose that in the absence of after-school care, a student would have cared for himself or herself after school until a parent arrived at 5 p.m. A correct response for this student would have involved reporting two responses: "I was by myself" and that "My Mom or Dad" provided care too, after regular school until 6 p.m. The authors' coding algorithm would have coded the student as being in parental care. Now, suppose that the child attended the after-school program, rather than being alone after school, and then met his or her mother or father at home at 5:15 p.m. The correct response would have involved reporting two responses again: that "an adult who is not my parent" and "my Mom or Dad" provided care after school. Again, the student would have been coded as being in parental care, because parental care was at the top of the hierarchy for resolving cases with multiple responses. The authors would have estimated no impact for this child.

A second problem with the measure is that it is based upon the care arrangement on at least three days of the prior week—essentially using the modal care arrangement, not the mean care arrangement. This is important given the inconsistent attendance patterns the authors reported. For example, suppose that the program led to a change in care arrangements two days every week—a sizeable impact, but not enough to change the mode. Instead of reporting a 40 percentage point impact (two out of five days per week), the reported impact would have been zero since the modal care arrangement the other three days per week would have been unchanged.

The researchers considered using time diaries, but decided against doing so because of

the time commitment required for participants. (The provocative results from the 21st CCLC evaluation suggest that future evaluations should reconsider this decision.) But even if we were restricted to the questions the evaluators asked, a more helpful measure would have been the percentage of days with *some* parental care after school, *some* time spent alone, and the other categories. (Such categories would not necessarily be mutually exclusive.)

Third, the impacts on after-school care were measured at the time of the follow-up, when many youth had stopped attending the program. The impacts may have been larger early on.

Finally, students may have been reluctant to report that they are alone after school. (Youth may be coached by their parents not to tell strangers when there is not an adult at home.) This may also have led to an understatement of the impact. For example, if only 20 percent of the children who are alone (or the parents who leave their children alone) are willing to admit it, the estimated impact would have been only 20 percent as large as the actual impact. The authors of the 21st CCLC evaluation reported that the rates of self-care in the sample were comparable to the rates reported in the National Survey of American Families. However, all such measures may have the same problem.

School Attendance

In the 21st CCLC middle school evaluation, after-school participants had fewer absences and less tardiness.[14] In the TASC evaluation, active participants had larger increases in school attendance relative to the baseline than non-participants. However, since participants in both of these evaluations had volunteered to spend even more time than required on school grounds, one might suspect that participants might have been more likely to attend school even without the program. The TASC impact estimate is particularly difficult to interpret, given the definition of "active participation." Children are presumably more likely to attend the after-school program if they attended regular school that day. As a result, any unmeasured factor that led to more regular attendance during regular school is likely to have led to more regular attendance in the after-school program.

Given the selection bias problem, the most credible impact estimated for this outcome comes from the 21st CCLC elementary school evaluation, which showed no impact on absences or tardiness. Moreover, neither the ESS nor the SFBI evaluation reported statistically significant impacts on student absences.[15]

[14] The 21st CCLC middle school evaluation suggested that the program led to an *increase* in the mean days attending after-school activities such as band, drama, clubs, and sports. In the elementary school evaluation, there were also large—but not statistically significant—positive impacts on band, drama, and music/art/dance lesson participation (and a statistically significant negative impact on club participation, such as Boy Scouts).

[15] On this point, the results in the ESS evaluation are somewhat puzzling. The authors found no significant impact on the proportion of students reporting "skipping school." But they did report a large negative impact on the proportion of students reporting that they "started skipping school." ESS Evaluation, p. 65.

Grade Point Average

In the 21st CCLC elementary evaluation, there was a statistically significant and positive impact on grades in social studies/history. The point estimates for math, English, and science grades were also positive, albeit non-significant. In the 21st CCLC middle school evaluation, the impacts on math grades were statistically significant, but they were not significant for English, science, social studies, or history. The SFBI evaluation reported no statistically significant impact on school grades.

Homework Completion

Despite the failure to find robust impacts on student grades, there was more consistent evidence that participation in after-school programs led students to do their homework more diligently. Although, in the 21st CCLC elementary and middle school evaluations, there did not seem to be any impact on the proportion of students self-reporting that they "often" or "always" do their homework (one might wonder, however, whether the youth might have an incentive to exaggerate), there was an increase in the proportion of teachers "agreeing" or "strongly agreeing" that the students completed the assignment to the teacher's satisfaction in the middle school evaluation. In the elementary school evaluation, there was an increase in the proportion of teachers "agreeing" or "strongly agreeing" that children "usually try hard" in reading or English.

In the ESS evaluation, there was a positive impact on the students' self-reports of paying attention in class. In the SFBI evaluation there was a positive impact on the students' self-reported level of effort. The TASC evaluation reported higher levels of school engagement among participants, but did not collect such data for non-participants.

Parental Participation

Improved parental engagement in school is not the primary goal of after-school programming. It is ironic, then, that parental participation in after-school events was the outcome with the most consistent positive impacts. In the 21st CCLC middle school evaluation, there were large impacts on parental attendance at open houses, parent-teacher events, and volunteer activity at the school. Given that active parents were also probably more likely to hear about, and be recruited into, the after-school programs, one might take such a result with a grain of salt. However, even in the 21st CCLC elementary school evaluation, parents were statistically significantly more likely to help youth with their homework and to attend after-school events. Given the importance of parental involvement in schooling, this is an encouraging result.

VI. Interpreting the Test Score Impact Estimates

With the increasing emphasis on test-based accountability systems in elementary and secondary schools, the impact of after-school programs on measurable academic achievement is,

inevitably, of special interest to policymakers and voters. With the passage of the No Child Left Behind Act of 2001, states are under increasing pressure to raise student performance for all groups—particularly disadvantaged minority youth.

None of the evaluations under review reported a statistically significant impact on test scores at the end of a single year of participation. In the 21st CCLC elementary school evaluation, there was no statistically significant impact on Stanford 9 reading test scores. Likewise, the SFBI evaluation reported no impacts of participation on math or reading test scores. The TASC evaluation failed to find any statistically significant impacts on reading or math scores in the first year of attendance.

The TASC evaluation did report modest impacts on math scores for those who participated in the second year and somewhat larger impacts for those who were "active participants" in the second and third years. However, one should be reluctant to rely too heavily on the math impacts in the second and third year of the TASC evaluation—particularly for the "active participants"—given the potential for selection bias for those remaining in the program over several years. Interestingly, even for these "active participants," there was no statistically significant impact on reading scores after three years. (The ESS evaluation did not report any impacts on standardized test scores.)

However, beyond reporting the magnitude of the impacts that were found, none of the evaluations considered the magnitude of the impact we should have expected from the type of intervention being evaluated. In the education field, it has become common practice to design evaluations to identify an impact of one-tenth to three-tenths of a standard deviation. The perception is that impacts of such magnitude are believed to be important for policy and smaller impacts would not be. But such arbitrary standards can lead researchers to set unrealistic expectations. Test performance varies in the population for a variety of reasons—such as family and neighborhood influences. The quality of the academic instruction to which children are exposed—in regular school or in an after-school program—may be difficult to detect amidst the variation attributable to these other factors.

Using the Stanford 9 Spring and Fall Scores to Infer the Effect of Classroom Instruction

Harcourt Educational Measurement, the publishers of the Stanford 9 achievement test, administered their tests to a sample of 250,000 youth in the spring of 1995 (during the period April 3 to April 28, 1995) and to 200,000 youth in the fall of 1995 (during the period September 18 to October 13, 1995). They published the mean scores and standard deviations for the samples of students taking the exams at different grade levels in the fall and spring.[16] The "scaled" scores are intended to allow comparisons of scores across different grades, even though the test items used at different grade levels did not fully overlap. (In other words, even though some of the math items are different for fourth and fifth grade students, the items have been weighted according to their difficulty and put on a similar scale.) The top panels of Figure 1 report the mean scaled scores in math and reading for those tested in the spring of first grade through the spring of sixth grade. Over the five grade levels between first grade and sixth grade,

[16] Harcourt Educational Measurement (1996), Tables N2 and N5.

the mean scaled scores for students differed by 151 points in math and 155 points in reading. The standard deviation in scaled scores within each grade averaged 38 points in math (ranging in a narrow band between 36.5 and 40.7) and 43 points in reading (ranging somewhat more from a maximum of 51.1 in second grade to a low of 38.1 in sixth grade). Therefore, if we were to assume that the performance of the students in a given grade in 1995 was a reasonable proxy for the performance of students in the previous grade in 1994, we would infer that test performance grew by approximately .7 standard deviations in reading and .8 standard deviations in math per grade level. But the rise in performance at different grade levels flattens out considerably after grade three. For instance, between the end of grades five and six, math and reading scores grew by .30 standard deviations in math and .25 standard deviations in reading. In other words, everything that happens to a student between the spring of fifth grade and the spring of sixth grade—a full year of classroom instruction, a year of conversations with parents at the dinner table, spending time with friends, and being exposed to cognitive stimuli of all kinds—is associated with .25 to .30 standard deviations gain in performance.

The growth in academic achievement between grade levels reflects more than the effect of classroom instruction; it includes the effect of all cognitive stimuli from family, friends and the environment. One way to sort out the specific contribution of classroom instruction would be to compare student test scores in the fall and spring of the same grade to the growth in scaled scores between the spring in one grade and the fall in the subsequent grade. The spring and fall testing sessions were roughly six months apart (168 days from midpoint to midpoint of the testing sessions). However, the number of days during which students were exposed to classroom instruction would have been very different between fall and spring and between spring and fall. Between the fall and spring testing, students would have spent all but several weeks in classroom instruction, but between the spring and fall testing, students will have spent much of their time on summer vacation. Suppose we were to assume that the flow of cognitive stimuli *outside of school* was relatively constant all year. If an hour of classroom instruction had the same effect on test scores as an hour of life experience, we would expect the growth in performance between fall and spring to be roughly equal to the growth in performance between spring and fall. On the other hand, if students gain more from their time in the classroom, we would expect performance to grow more from fall to spring. Comparing the growth between fall and spring and between spring and fall gives us a way to identify the effect of classroom instruction as distinct from the general effect of life experience and maturation.

The bottom panels of Figure 1 report the fall to spring comparisons from one academic year to the next and the spring to fall comparisons within the same academic year of scale scores in math as well as reading. There are several facts worth noting. First, in grades two and three, the fall to spring gains were generally larger than the spring to fall gains in both math and reading. This is reassuring, because it implies that children learn more from their time spent in school than their time spent on summer vacation. Second, the difference between the two lines is larger in math than in reading, implying that instructional time makes more of a difference for math than for reading. Third, both lines decline and the gap between the two lines diminishes with age—implying that the rate at which students learn declines and the differential impact of instructional time lessens as students age. (Indeed, by fourth grade the lines have crossed for reading, implying that students improve about as much between April and October as between October and April.)

The average difference in scores between the spring and fall and the fall and spring in the subsequent year was .11 standard deviation. (The average difference in distance between the two lines in the bottom panel of Figure 1 was .11 standard deviations.) Students do typically learn more between fall and spring than between spring and fall. But what does this imply about the magnitude of the gain due to instruction? Assuming 36 full weeks of classroom instruction in a typical school year (180 day school year), and assuming that youth attend schools all but two weeks between the fall and spring testing sessions, students will have attended 26 weeks of classes between the fall and spring and would have received ten weeks of instruction between the spring and fall.[17] In other words, students will have attended .72 academic years of instruction between fall and spring and .28 academic years of instruction between spring and fall. Dividing the .11 standard deviation differential in performance improvement by the .44 differential in years of academic training received implies that an academic year of instruction is associated with a .248 increase in academic performance.

Using Cut-offs for Age of Entry into Kindergarten

Another way to identify the effect of classroom instruction would be to use the age of school entry limits by school districts to identify when students start school. In the Los Angeles Unified School District (LAUSD), students must celebrate their fifth birthday on or before December 2 of a school year in order to start kindergarten that year.[18] Students born on December 2 are allowed to enroll in the fall of the year they reach age five, but students born on December 3 are required to wait a year. This provides a convenient natural experiment. When we see them several years later, the youths born on December 2 and December 3 will have had very similar amounts of total life experience. But because of the school district's enrollment policy, they will have spent very different shares of that time in a classroom.

Of course, some parents of those born on December 2 will decide to wait and have their children enroll in the following year anyway. These are probably not a random sample of the youth born on that day—the most mature of those with a December 2 birthday will be allowed to start the year they turn five and the less mature children will be held back. As a result, one would not want to compare the fifth grade students with December 2 birthdays to the fourth grade students with December 3 birthdays, because only the most mature December 2 birthday students will actually be in fifth grade. This would probably overstate the impact of a year of classroom training. However, if one were to focus on the difference in the average score of all students—in whatever grade they may be enrolled—born on December 2 and December 3 of the same year, this would provide us with a measure of the amount of learning attributable to the difference in amount of time the two groups were in a classroom. (Of course, the proposed calculation relies heavily on the methods used to put the results of different grade-level tests on the same scale.)

[17] This oversimplifies greatly, since there are some school districts, such as Los Angeles, in which students are attending school on a year-round calendar.

[18] Memorandum from Assistant Superintendent Maria Reza, "Memorandum No. Z-20 (Rev.), Entrance Ages; Kindergarten Continuation and Verification of Birthday", Los Angeles Unified School District, June 15, 2001.

Figure 2 reports the mean grade students are attending as well as mean reading and math scores in spring 2002, by single day of birth. (There were 160 kids with valid test scores with the same date of birth in the LAUSD.) The vertical lines are drawn at December 2, 1991, 1992, 1993, and 1994, the cut-off dates. The very sharp discontinuities in the mean grade students are attending at the December 2 cut-off dates suggest that the cut-off date is binding for most youth. In the spring of 2002, students born on December 2 in any year had completed about .8 grade levels more than students born on December 3.

The other two panels in Figure 2 report mean scaled scores in math and reading, also arranged by single day of birth. Two facts should be apparent in these other panels. First, age clearly is positively related to test performance even within grade. Presumably, this reflects the value of the accumulation of cognitive stimuli outside of school as well as maturation. Second, there are discontinuities in test performance at the date of birth cut-offs in reading and math scores as well. These are particularly interesting because they reflect the value of the extra .8 years of educational attainment that those born on December 2 enjoy over those born on December 3. Each of these two groups of youth has accumulated very similar amounts of total life experience, but has enjoyed very different amounts of time in the classroom. Taking the discontinuities in the mean scaled scores and dividing by the discontinuities in grades completed (analogous to an instrumental variables estimator, using the cut-off dates as an instrument) implies a payoff per year of schooling of .05 to .17 standard deviations in reading (the discontinuities vary somewhat with age) and from 0 to .15 standard deviations in math. (See Appendix Table 1 for the results.)

Neal and Johnson (1996) use a related strategy, comparing Armed Forces Qualification Test scores (AFQT) by quarter of birth, using data in the National Longitudinal Survey of Youth (NLSY). Those born late in any calendar year are likely to have completed less schooling at any survey date in late adolescence than those born earlier in the year, because of the cut-offs determining age at school entry. Essentially dividing the difference in test scores for youth born in different quarters by the difference in years of schooling completed, Neal and Johnson report that a year of schooling raises AFQT scores for young men and women by .25 standard deviations. In a more recent paper using the same data but a different identification strategy, Hansen, Heckman, and Mullen (2003) report similar results. Cascio and Lewis (2003) also use the NLSY, but they explicitly incorporate the age at school entry policies in the various states. They report that a year of educational attainment is associated with a smaller .10 standard deviation improvement in AFQT scores.

Of course, the impacts are likely to vary with the type of skills being tested. Other evidence suggests that reading test scores may be slower to respond to instructional interventions than math scores. Reading is more likely to be driven by family background—the extent to which parents have read to their children outside of school—than math test scores, which depend more upon the specific topics the student has covered in school. For example, over the last two decades, math test scores have risen for nearly every age and race group in the National Assessment of Educational Progress.[19] There has been little or no progress in raising reading scores.

[19] For the national reading trends, see http://nces.ed.gov/nationsreportcard/reading/trendsnational.asp. For the national math trends, see http://nces.ed.gov/nationsreportcard/mathematics/trendsnational.asp.

Consistent with the hypothesis that instructional quality matters less for reading than for math, there is evidence that schools differ much less in the extent to which they boost reading achievement as compared with math achievement. When studying the variability in student gains in reading between schools in North Carolina (and adjusting for student demographics and parental education), Kane and Staiger (2001) report much less variation between schools in mean student gains on that state's reading test than on the math test. In Texas, Hanushek, Rivkin, and Kain (1998) also report much less between-school variation in mean student achievement increases in reading than in math. Similarly, summer school interventions have typically been found to have larger impacts on math achievement than on reading achievement. (For a meta-analysis of these results, see Cooper et. al. [2000] and Cooper [2001].)

What is a Reasonable Expectation of Test Impacts for After-School Programs?

There are at least two ways to think about forming expectations for the achievement impact of after-school programs. One is to start with our best estimates of the impact of classroom instruction and ask what they would imply about the impact of an additional hour of academic instruction during an after-school program, even if one participated five days per week. Suppose we were to start with the estimate that an entire year of classroom instruction is associated with a quarter of a standard deviation improvement in achievement test scores. This is consistent with the analysis in Neal and Johnson (1996) and in the analysis of Stanford 9 scaled scores above. If the typical regular school day were to include five hours of course instruction (to account for recess and movement between classes), providing an extra hour per day in academic instruction would lead us to expect roughly .05 student-level standard deviations ([1 hour/5 hours] x .26). (If we had used the other estimates above—such as those derived from the date-of-birth cut-offs for LAUSD—we would have generated even more modest expectations.)

One reason for the limited impact of the after-school programs in the evaluations is that they failed to achieve 100 percent participation. But note that the above calculation assumed that all participants attended every day and that non-participants received no classroom instruction after regular school. If we were to account for anything less than full-participation or if we were to recognize that some members of the control group may also have access to after-school programs, our expectation would be even more modest.

A second way to think about forming such an expectation would be to start with the dollar value of test score gains to students and society later in life and then ask how large an impact on student test scores would be required to fully pay off the cost of after-school care provided. Kane and Staiger (2002) use two recent estimates of the relationship between test performance and earnings of young adults to construct such a measure. (Krueger [2002] does a similar calculation.) Calculating the present value of the increase in lifetime earnings when a child is in grade four, the lifetime earnings increase associated with a one-standard deviation improvement in test performance is between $90,000 and $210,000. Seven one-hundredths of a standard deviation improvement in test scores would be worth $6,300 to $14,700 in present value per youth, more than the cost of a slot in the average after-school program.

In the 21st CCLC evaluation, the reading test score impact estimate was .1 percentile points (with a mean score for the comparison group at the 34.1 percentile of the national distribution). The point estimate is obviously quite close to zero. However, the estimate also has a standard error of 2 percentile points.[20] Assuming that the norm sample of test scores is normal, the 95 percent confidence interval would extend roughly .2 national standard deviations. In other words, while the confidence interval for the impact estimate includes zero, it also includes most of the impacts we might reasonably have expected.

The 21st CCLC evaluation was designed to be able to identify a .20 impact—three to four times the impact that would reasonably have been expected even with full participation. Moreover, the elementary school evaluation collected only Stanford 9 reading test scores, not math test scores. Both factors made it very unlikely that the evaluation would have led researchers to be able to reject the null hypothesis that the programs had zero impact on performance.

What does the existing evidence suggest about the impact of after-school programs on student achievement? We have learned that after-school programs do not lead to extraordinarily large increases in reading achievement. However, we do not know whether after-school programs may be having more moderate, but nevertheless worthwhile, impacts on other academic performance measures.

VII. Conclusion

After-school programs have generated a lot of interest—as a way to make better use of public buildings in after-school hours, to improve student safety and, most recently, to improve academic achievement. However, attendance in such programs is usually sporadic. Participating students typically attend one to two days per week. Moreover, not all of the participating students would have been unsupervised by an adult in the absence of the program. Indeed, the 21st CCLC evaluation suggests that many of the participating students would have been at home with a parent. If future results confirm such findings, they would raise the bar for after-school programs: not only must they offer a safe environment for youth, they must also be more worthwhile than a couple of additional hours at home after school. Programs need to do a better job of identifying the youth who are currently home alone or wandering the streets after school, and persuade them or their parents to participate.

Despite the low attendance rates, there are some promising results suggesting that parents are more likely to participate in schools and that students are more likely to pursue their homework more diligently. The impacts on students' grade point averages were less consistent, but are also somewhat encouraging.

The above discussion offers a number of lessons for future evaluation designs. First, given that participants choose to attend and that non-participants face the same options and do not, the interpretation of quasi-experimental estimates using participants and non-participants

[20] I inferred this using the p-value of .96 reported in Dynarski, et. al. (2003), Table IV.5.

attending the same schools will always be problematic. An alternative approach would be to exploit differences in the timing of introduction of after-school programs between schools. (In either case, the collection of baseline outcome data that would allow us to test whether regression-adjustment eliminates pre-treatment differences between participants and non-participants is important.) Still, random assignment studies are likely to yield the most plausible impact estimates in this field.

Second, when evaluating academic achievement impacts, future evaluations should consider administering both math and reading tests, preferably at baseline and at follow-up. School systems often have incomplete testing data for students. Moreover, given the modest increment in cost—particularly in light of the considerable investment involved in conducting any experimental evaluation—the decision to exclude math achievement and administer only a reading test in the 21st CCLC evaluation was regrettable.

Third, given the importance of after-school care arrangements in the rationale for after-school programs, future evaluations should invest more resources in measuring and comparing the after-school care arrangements used by treatment and comparison groups. The surprisingly small impacts on adult supervision after school reported in the 21st CCLC warrant further study.

Finally, given the time devoted to academic activities and the magnitude of resources devoted to after-school programs, it seems unrealistic to expect large impacts on academic achievement (such as .2 standard deviations). Unless an hour spent in an after-school program is extraordinarily more productive than an hour spent in regular instruction, such large impacts seem unlikely. Future evaluation work needs to be more explicit in identifying the magnitude of impact that might reasonably be expected—given the projected cost of the program or given external estimates of the impact of a given amount of time in academic instruction. Either approach is likely to suggest impacts much smaller than .2 standard deviations. Unless the resources are available for collecting the size of sample required to identify academic achievement impacts of a small magnitude, future evaluations run the risk of passing over worthwhile interventions. As an alternative to collecting large samples, such evaluations should concentrate resources on measuring intermediate outcomes—such as homework completion or parental participation—for which impacts may be more readily discernible.

References

Bissell, Joan S., Christopher Cross, Karen Mapp, Elizabeth Reisner, Deborah Lowe Vandell, Constancia Warren, and Richard Weissbourd "Statement Released by Members of the Scientific Advisory Board for the 21st Century Community Learning Center Evaluation" May 10, 2003.

Cascio, Elizabeth and Ethan Lewis, "Regression Discontinuity Evidence on the Effect of Schooling on the AFQT" University of California-Berkeley Working Paper, June 2003.

Cooper, Harris, Kelly Charlton, Jeff Valentine, and Laura Muhlenbruck *Making the Most of Summer School* (Malden, MA: Blackwell, Monographs Series of the Society for Research in Child Development, 2000).

Cooper, Harris "Summer School: Research-Based Recommendations for Policymakers" SERVE Policy Brief, 2001.

Dynarski, Mark et. al., "When Schools Stay Open Late: The National Evaluation of the 21st Century Community Learning Centers Program" (Washington, DC: U.S. Department of Education, Office of the Under Secretary, January 2003).

Grossman, Jean Baldwin et. al., "Multiple Choices After School: Findings from the Extended-Service Schools Initiatives" (New York, NY: Public/Private Ventures and MDRC, June 2002).

Hansen, Karsten, James J. Heckman, and Kathleen Mullen "The Effect of Schooling and Ability on Achievement Test Scores" National Bureau of Economic Research Working Paper No. 9881, July 2003.

Hanushek, Eric, John Kain, and Steve Rivkin "Teachers, Schools and Academic Achievement" National Bureau of Economic Research Working Paper No. 6691, August 1998.

Harcourt Educational Measurement, Stanford Achievement Test Series, Ninth Edition, Technical Data Report, (San Antonio, TX: Harcourt, 1996).

Jacobson, Linda "After-School Report Called Into Question" Education Week (May 21, 2003) Vol. 22, No. No. 37, p. 1, 15.

Kane, Thomas J. and Douglas O. Staiger "Improving School Accountability Measures" National Bureau of Economic Research Working Paper No. 8156, March 2001.

Kane, Thomas J. and Douglas O. Staiger "The Promise and Pitfalls of Using Imprecise School Accountability Measures" Journal of Economic Perspectives (Fall, 2002), Vol. 16, No. 4, pp. 91-114.

Krueger, Alan B. "Economic Considerations and Class Size" National Bureau of Economic Research Working Paper No. 8875, April 2002.

Mahoney, Joseph L. and Edward F. Zigler "The National Evaluation of the 21st Century Community Learning Centers: A Critical Analysis of First-Year Findings" Working Paper, Department of Psychology, Yale University, July 3, 2003.

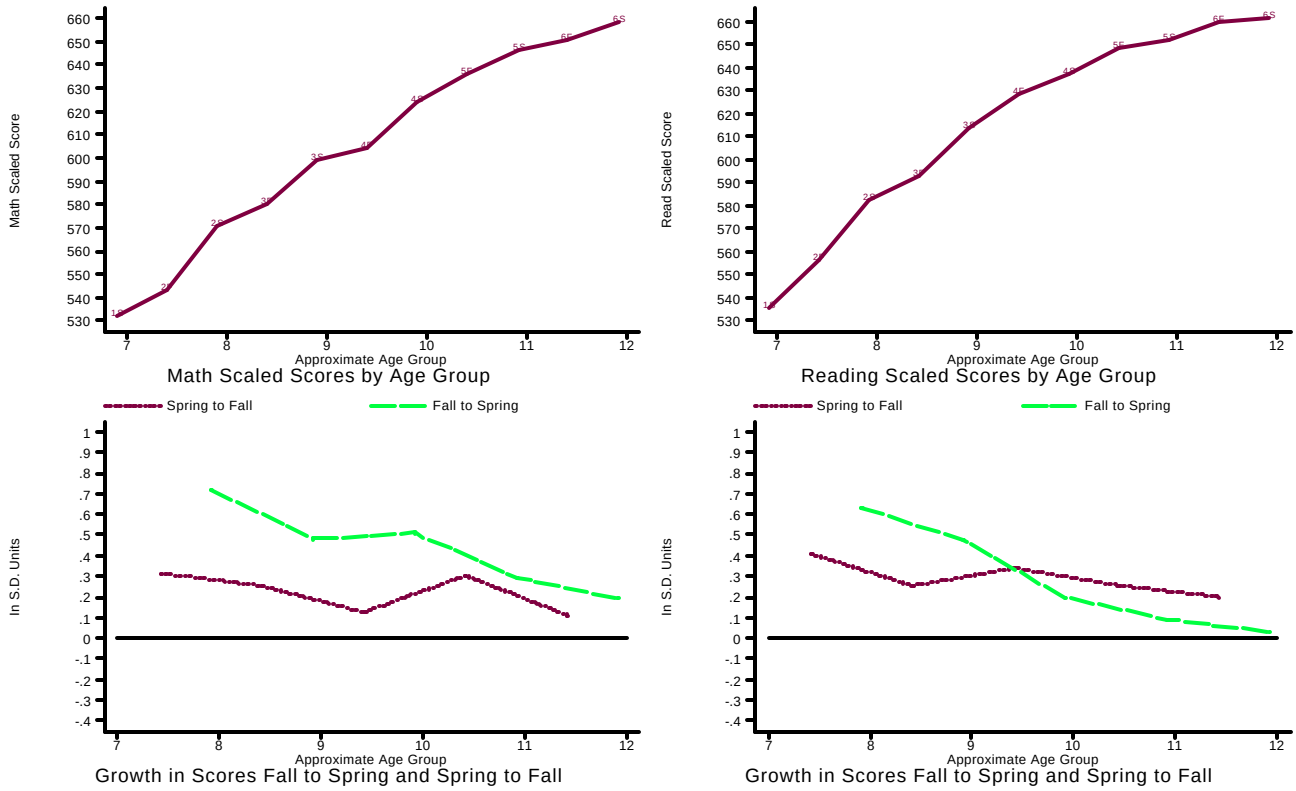
Maxfield, Myles, Allen Schirm, and Nuria Rodriguez-Planas, "The Quantum Opportunities Program Demonstration: Implementation and Short-Term Impacts," (Washington, DC: Mathematica Policy Research, December 2002) (U.S. DOL Contract No. K-5547-5-00-80-30).

Reisner, Elizabeth R., Christina Russell, Megan Welsh, Jennifer Birmingham, and Richard White "Supporting Quality and Scale in After-School Services to Urban Youth: Evaluation of Program Implementation and Student Engagement in the TASC After-School Program's Third Year" Policy Studies Associates report to The After-School Corporation, March 29, 2002.

Walker, Karen et. al., "San Francisco Beacons Initiative Final Report", Draft, May 2003.

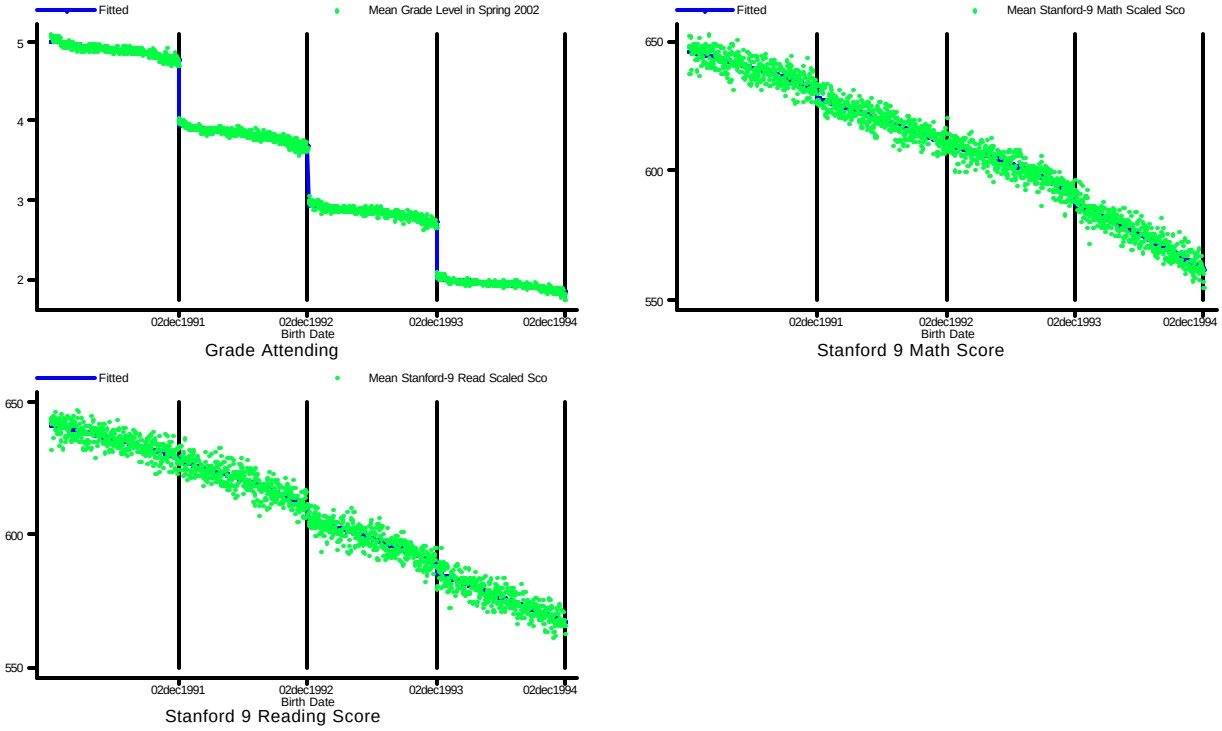
Welsh, Megan, Christina Russell, Imeh Williams, Elizabeth Reisner and Richard White "Promoting Learning and School Attendance through After-School Programs" (Washington, DC: Policy Studies Associates, October 31, 2002).

Figure 1.



The Growth in Scaled Scores Across Semesters and Grades

Figure 2:



(by Day of Birth)

Mean Grade Attending, Math and Reading Scores in Spring 2002

Table 1. After-School Program Evaluations Included in This Synthesis

	21st CCLC (Elementary)	21st CCLC (Middle Schools)	TASC	ESS	SFBI
Evaluator	Mathematica Policy Research and Decision Information Resources	Mathematica Policy Research and Decision Information Resources	Policy Studies Associates	Public/Private Ventures and MDRC	Public/Private Ventures
Sites	11 Oversubscribed elementary school centers	Representative sample of 46 middle school centers nationally	All 96 projects first funded in 1998-99 and 1999-00 school years	10 Schools in 6 cities	3 middle school centers in S.F. (Also some data for 1 elementary center and 1 high school center.)
Dosage Evaluated	1 year	1 year	Up to 3 years	1 year	Up to 3 years
Comparison Group	Randomly assigned applicants	Non-participants in same schools	Non-participants in same schools	Non-participants in same schools	Non-participants in host schools
Data Collection	Baseline: Student and parent survey Follow-up: Student, parent, and teacher survey Participation: Daily logs School Records: Demographics, free lunch status, attendance, and grades Test Scores: Reading score baseline and follow-up (administered by research team and schools) Program Data: Nature of activities, staffing, costs.	Baseline: Student survey Follow-up: Student, parent, and teacher survey Participation: Daily logs School Records: Demographics, free lunch status, attendance, and grades Test Scores: Incomplete baseline and follow-up (administered by schools) Program Data: Nature of activities, staffing, costs.	Baseline: School records and test scores Follow-up: Student, parent, and school and program staff surveys. Participation: Daily logs School Records: Demographics, free lunch status, and attendance Test Scores: Baseline and follow-up (administered by schools) Program Data: Nature of activities, staffing, costs.	Baseline: Student survey Follow-up: Student survey Participation: Daily logs Program Data: Nature of activities, staffing, costs.	Baseline: Student survey Follow-up: Student survey Participation: Daily logs School records: Demographics, free lunch status, attendance, test scores, and grades Program Data: Nature of activities, staffing, costs.
Outcomes Studied	Participation, after-school care arrangements, subjective feeling of safety, grades, reading test scores, school attendance, homework completion, parental involvement in school	Participation, after-school care arrangements, subjective feeling of safety, grades, school attendance, homework completion, parental involvement in school	Participation, school attendance, test scores	Participation, after-school activities	Participation, after-school activities, school attendance, grades, math and reading test scores.

Table 2. Program Descriptions and Participation Rates

	21st CCLC (Elementary)	21st CCLC (Middle Schools)	TASC	ESS	SFBI
Definition of "Ever Participant"	Parents applied to participate and student attended at least 1 day	Attended 3+ days in first four weeks of program	Participants attended at least one day "Active Participants" attended 60 or more days or 60 percent of scheduled days	Participated at least one day during school year	Participated at least once during the school year or summer session
Percent of Host School Youth Participating	Percent applying not reported.	12%	39%	>50%	47%
Average Daily Attendance Rate (by participants)	37%	18%	78 % Elementary 57 % Middle Schools	47% Grade 1-3 35% Grade 4-5 23% Grade 6-8	30% Middle School (inferred from 1.5 day/week dosage)
Avg. days per week	1.9	.9	3.9 Elementary 2.9 Middle Schools (% att. x 5 days)	2.4 Grade 1-3 1.8 Grade 4-5 1.2 Grade 6-8	1.5
Participation Requirement	No	No	Strongly Encouraged Full-time	No	No
Continuation Rate between years	n.a.	n.a.	46-48%	n.a.	n.a.
Center Staffing	35 % teachers	40 % teachers	24% teachers and other school staff	n.a.	0-15% teachers at school
Cost per day	n.a.	n.a.	$6.76 per slot (excludes costs of services provided by others and capital costs)	$15 per slot (includes cost of services provided by others, but excludes capital costs)	$27 per slot (includes administrative and support costs, but excludes capital costs)

Table 3. Summary of Impact Estimates: 21st CCLC and TASC

Type of Outcome:	21st CCLC (Elementary)	21st CCLC (Middle Schools)	TASC
Evaluation Strategy	Random assignment with regression-adjustment	Regression-adjusted	Regression-adjusted
Regressors	Race/Ethnicity, gender, student baseline test scores, baseline attendance, household socioeconomic status	Race/Ethnicity, English language learner status, gender, age, baseline grades, family income, mother college degree, public assistance receipt, household structure, disciplinary record, absences, times tardy, prior retention	Baseline score, free lunch status, gender, school poverty rate, grade, race/ethnicity, English language learner status, special education, recent immigrant
After-School Care Arrangement	Substituted non-parent adult care for parent care *and* sibling care Large but non-significant impacts on band, drama, music/art/dance lessons Negative impact on club attendance No impact on student perceived safety after school until 6 p.m.	Substituted non-parent adult care for parent care *and* sibling care and reduced presence in someone else's home after school Increased mean days of tutoring, band, drama, clubs, sports No impact on student perceived safety after school until 6 p.m.	n.a.
School Attendance	No impact on absences, tardiness	Fewer absences, less tardiness	"Active Participants" increased attendance more than non-participants
Homework Completion	No homework effect, but increase in teacher reports that students "usually try hard" in reading or English	No homework effect but more likely to complete to teacher's satisfaction	n.a.
Parental Participation	Increase in parental help with homework and attendance at after-school events	Large impacts on attendance at open houses, parent-teacher events, volunteering	n.a.
Grades	Increased social studies/history grades, but impacts on math, English, and science were not statistically significant.	Increased math grades, but not English, science, social studies, or history	n.a.
Test Scores	No impact on reading test scores	n.a.	*Participants vs. non-participants: (in standard deviation. units):* Math: -.08 ** After 1 year .12 ** After 2 years n.s. After 3 years Reading: n.s. *"Active" participants every year vs. non-participants:* Math: n.s. After 1 year .17** After 2 years .17** After 3 years Reading: n.s. **=Significant at 5 percent level.

Table 4. Summary of Impact Estimates: ESS and SFBI

Type of Outcome:	ESS	SFBI
Evaluation Strategy	Regression	Regression
Regressors	Gender, race/ethnicity, low-income status, single-parent household status, parent high school dropout, parent high school graduate, number of prior moves, attendance at religious institution, self-reported grades, school dummies, and baseline measures of school engagement and parent-child relationships	Gender, race/ethnicity, free-lunch status, grade, site dummies, baseline GPA, baseline test scores, baseline absences, set of baseline developmental and well-being measures
After-School Care Arrangement	n.a.	n.a.
School Attendance	No significant impact on student reports of skipping school[21]	No significant impact
Homework Completion/ Level of Effort	Positive impact on student self-report of paying attention in class	Positive impact on student self-reported level of effort
Parental Participation in School Activities	n.a.	n.a.
Grades	n.a.	No significant impact
Test Scores	n.a.	No significant impact

[21] No significant impact on student report of skipping school, but large negative impact on the proportion of kids reporting that they "started skipping school." ESS Evaluation, p. 65.

Appendix Table 1:
Weighted Least Squares Regression Coefficients From Regressing Grade Level and Stanford 9 Scores on Days Since Birth and Birthdate Cut-offs for School Entry

(Standard errors reported in parentheses.)

	Dependent Variable:		
Independent Variables:	Grade Level	Reading Score	Math Score
Days Since Birth/100	0.0356	4.9682	7.6080
(until June 1 2002)	(0.0028)	(0.2613)	(0.2712)
Days Since Birth^2	0.0034	0.0206	-0.2355
	(0.0005)	(0.0429)	(0.0445)
Days Since Birth^3	-0.0001	-0.0042	0.0049
	(0.00002)	(0.0019)	(0.0020)
Born Dec. 2, 1991	0.8278	1.5116	4.4639
or before	(0.0049)	(0.4595)	(0.4769)
Born Dec. 2, 1992	0.7432	5.3146	-0.6044
or before	(0.0052)	(0.4912)	(0.5099)
Born Dec. 2, 1993	0.7106	3.7729	4.4106
or before	(0.0048)	(0.4563)	(0.4736)
Constant	1.861	567.228	562.147
	(0.004)	(0.371)	(0.385)
Observations	1461	1461	1461
R-squared	0.99	0.98	0.98

Note: The above were estimated using mean grade level and mean scaled score by single day of birth in the Los Angeles Unified School District (LAUSD). The sample consisted of those born between December 3, 1990, and December 2, 1994. There were approximately 180 youth born on each day during the period enrolled in LAUSD. Outcomes were measured in the spring of 2002, when most of the sample members would have been between grades one and five. The coefficients were estimated by weighted least squares, using the square root of the sample size on each day of birth as the weight. The second and third rows contain regression coefficients for days since birth squared and days since birth cubed, respectively.